

Advanced Use of NCBI Resources

David Wheeler, Ph.D.

National Center for Biotechnology
Information

National Institutes of Health

Genomics and Proteomics
in Kidney and Urologic
Diseases

July 10, 2001

Advanced Use of NCBI Resources

Start Here:

<http://www.ncbi.nlm.nih.gov/>

Finding a New Gene

Finding a 3D Structural Model

Genomic Comparisons

Sifting Through Shotgun Sequence

Searching for an Unannotated Gene within the Human Genome

Evaluation of the Implied
Protein Sequence

Using LocusLink to Find a Probe Sequence

NCBI LocusLink

PubMed Entrez BLAST OMIM Taxonomy Structure

Search: LocusLink Display: Brief Organism: Human

Query: alpha oxoglutarate dehydrogenase Go Clear

View Loci Save Loci

LocusLink Home Help

2 loci found

LocusID	Org	Symbol	Description	Position	Links
<input type="checkbox"/> 55753	Hs	FLJ10851	hypothetical protein FLJ10851	10	P R G P H U V
<input type="checkbox"/> 4967	Hs	OGDH	oxoglutarate dehydrogenase (lipoamide)	7p14-p13	P O R G P H U V

The Protein RefSeq is the more Sensitive Probe

LocusLink Home

OGDH Index:

[Top of Page](#)
[Nomenclature](#)
[Overview](#)
[Function](#)
[Relationships](#)
[Map](#)
[RefSeq](#)
[GenBank](#)
[Links](#)

LocusLink:

[Collaborators](#)
[Download](#)
[FAQ](#)
[Help](#)
[Statistics](#)

RefSeq:

[About](#)
[Download](#)
[FAQ](#)
[Statistics](#)

NCBI Reference Sequences (RefSeq) ?

Category: **PROVISIONAL**

mRNA: [NM_002541](#)

Protein: [NP_002532](#) oxoglutarate dehydrogenase **BL**
(lipoamide)

Domains: [Dehydrogenase E1 component](#) score: 792

GenBank Source: [D10523](#)

Category: **NCBI Genome Annotation**

Genomic Contig: [NT_025782](#) **SV MV**

Evidence: supported by alignment with
mRNA

Model mRNA: [XM_004889](#)


Model Protein: [XP_004889](#) **BL**

Domains: [Dehydrogenase E1 component](#) score: 752

GenBank Sequences ?

Nucleotide	Type	Protein	
D32064	g	BAA06836	BL
BC004964	m	AAH04964	BL
D10523	m	BAA01393	BL

A tblastn Search using Human Genome BLAST



CGCTCAGGATAGAGACTTCGCGCGCTAGAGGATCGGATCCCCGGCGCATATTATATAGCTCGATCGATC
TTCTCTATATATAGCTCGCTAGAGGATCGGATCCCCGGCGCATATTATATAGCTCGATCGATC
Genome Sequencing
CCCCATCAGCAGCTAGAGGATAGAGGATCGGATCCCCGGCGCATATTATATAGCTCGATCGATC
CACAGACTGCATACGCATACGTCAGCTATACTTACTAACCAATTCGGGAGAGGGCGCGCGGATCGGC
GCAG

Chrs: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y

Search for

[Human Genome Sequencing](#)
[BLAST Home Page](#)
[BLAST overview](#)
[BLAST FAQs](#)
[BLAST news](#)
[BLAST manual](#)

BLAST the Human genome

Compare your query sequence to the working draft sequence of the human genome or its mRNA and protein products.

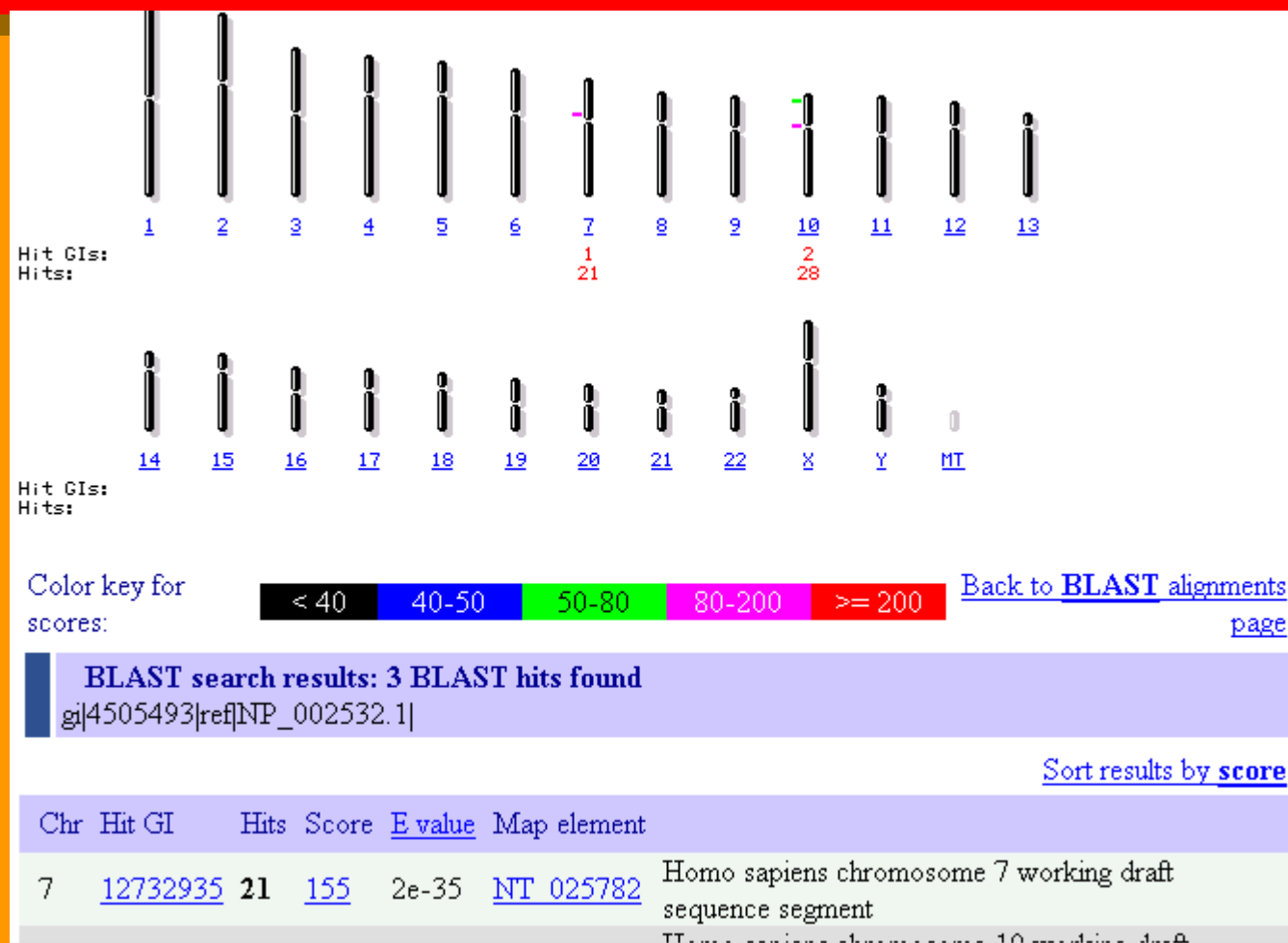
Database: Program:

Enter an accession, gi, or a sequence in FASTA format:

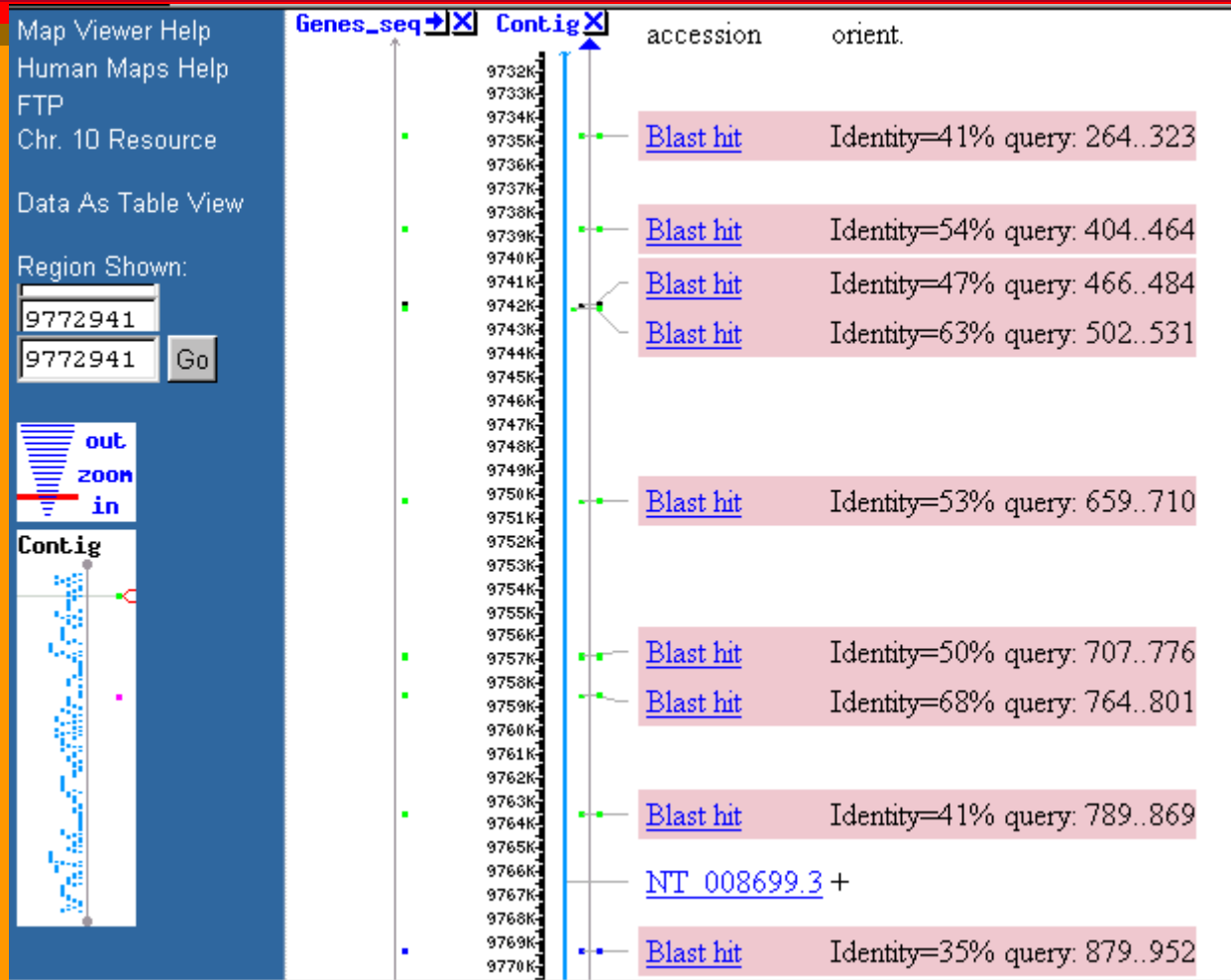
Optional parameters

[Expect](#) [Filter](#) [Descriptions](#) [Alignments](#)

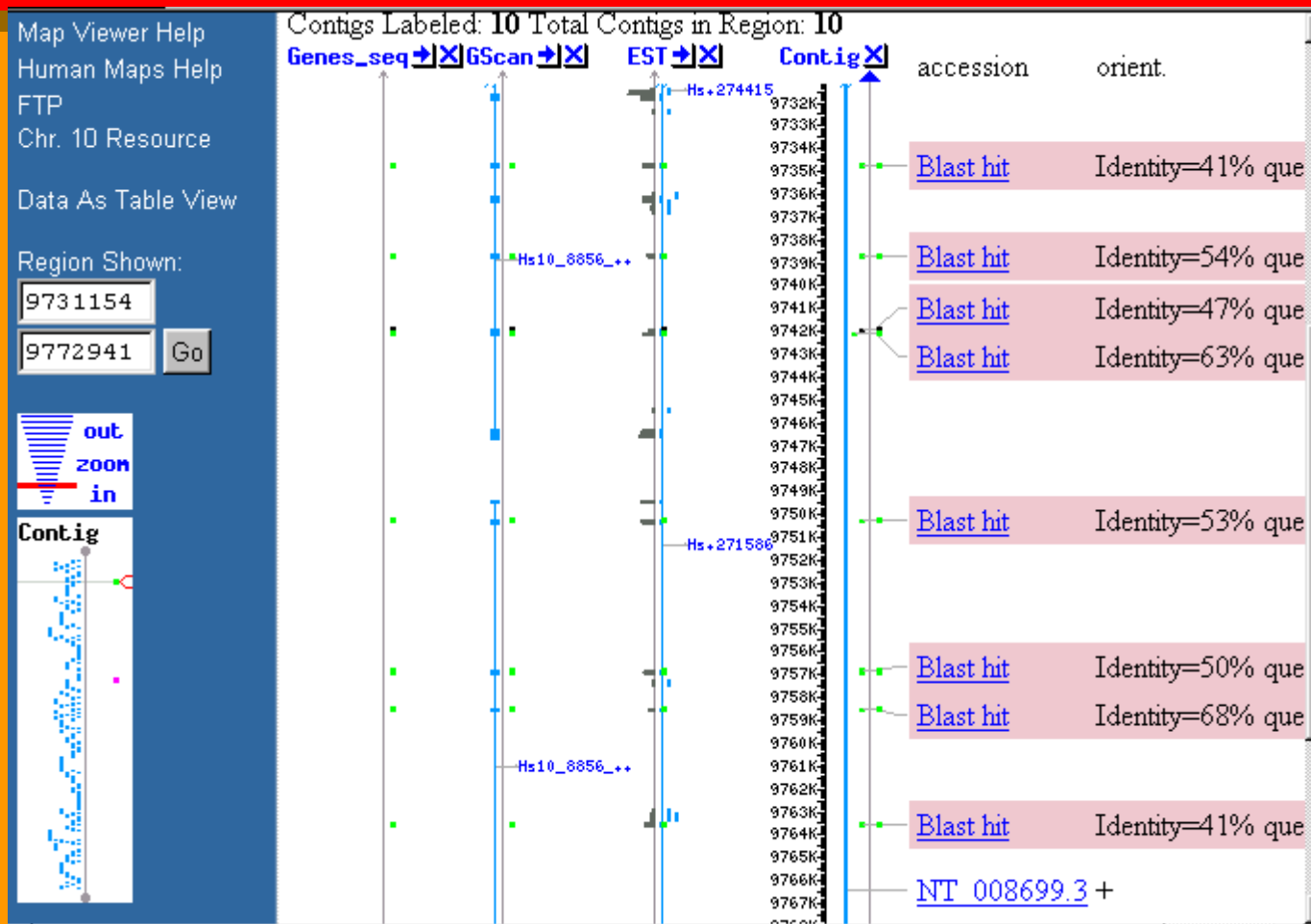
Three Hits



BLAST Hits but no RefSeq Alignments



GenomeScan Gene Models and EST Alignments



Some Information From Unigene Cluster 271586

Home Page

Frequently Asked
Questions

Query Tips

Library Differential
Display

Download
UniGene

UniGene
Homo sapiens

Home Page

Release Statistics

Library Report

Library Browser

SEE ALSO

LocusLink: [55526](#)

HomoloGene: [Hs.271586](#)

SELECTED MODEL ORGANISM PROTEIN SIMILARITIES

organism, protein and percent identity and length of aligned region

H.sapiens: [pir.T50617](#) - T50617 hypothetical protein 100 % / 539 aa
DKFZp762M115.1

M.musculus: [prf2001488A](#) - 2-OXOGLUTARATE 48 % / 128 aa
DEHYDROGENASE E1 COMPONENT

C.elegans: [pir.T28034](#) - T28034 hypothetical protein 50 % / 897 aa
ZK836.2 - Caenorhabditis elegans

S.cerevisiae: [sp.P20967](#) - ODO1_YEAST 38 % / 859 aa
2-OXOGLUTARATE DEHYDROGENASE E1
COMPONENT, MITOCHONDRIAL
PRECURSOR (ALPHA-KETOGLUTARATE D

E.coli: [pir.DEECOG](#) - DEECOG oxoglutarate 40 % / 856 aa
dehydrogenase (lipoamide) (EC 1.2.4.2) -
Escherichia coli

EXPRESSION INFORMATION

cDNA sources: Brain, Breast, Esophagus, Head and neck, Kidney, Tonsil,

Two GenomeScan Predicted Peptides

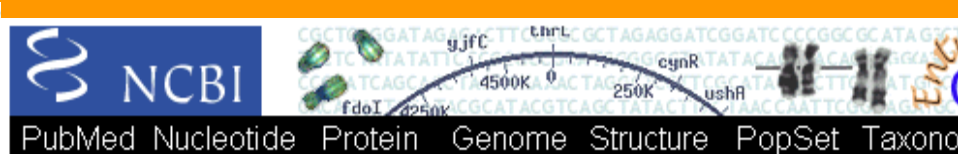


[Homo sapiens](#) Map View

GenomeScan model Hs10_8856_22_1_1

CDS:

```
>lc1|Hs10_8856_22_1_1
MASATAAAARRGLGRALPLLWRGYQTERGVYGYRPRKPESREPQGALERPPVDHGLARLV
TVYCEHGHKAAKINPLFTGQALLENVPEIQALVQTLQGPFFHTAGLLNMGKEEASLEEVLV
YLNQIYCGQISIETSQLQSODEKDWFAKRFEELQKETFTTEERKHL SKLMLESQEFDFHL
ATKFSTVKRYGGEGAESMMGFFHELLKMSAYSGITDVIIGMPHRGRNLNLTGLLQFPPEL
MFRKMRGLSEFPENFSATGDVLSHLTSSVDLYFGAHHPLHVTMLPNPSHLEAVNPVAVGK
TRGRQOSRQDGDYSPDNSAQPGDRVICLQVHGDAFCGQGVIPETFTLSNLPHFRIIGSV
HLIVMNQLGYTTTPAERGRSSLYCSDIGKLVGCAIIHVNGDSPEEVWGHNELDEFFYTNPI
MYKII RARKSIPDTYAEHLIAGGLMTQEEVSEIKSSYYAKLNDHLNMAHYRPPALNLQA
HWQGLAQPEAQITTWTSTGVPLDLLRFVGMKSVEVPRELQMHSLLKTHVQVGSLOMAGYC
FSFLLSKGADVGLVFSV
```



[Homo sapiens](#) Map View

GenomeScan model Hs10_8856_22_1_2

CDS:

```
>lc1|Hs10_8856_22_1_2
MMDGIKLDWATAEALALGSLLAQGFNVRLSGQDVGRGTFSQRHAIVVCQETDDTYIPLNH
MDPNQKGFLEVSNSPLSEEAVLGFYEGMSIESPKLLPLWEAQFGDFFNGAQIIFDTFISG
GEAKWLLQSGIVILLPHGYDGDAGPDHSSCRIERFLQAAVSTLQEMAPGTTFNPNVIGDSSV
DPKKVKTLVFCSGKHFYSLVKQRESLGAKKHDFAIIRVEELCPFFPLDSLQOEMSKYKHVK
DHIWSQEEPQNMGPWSFVSPRFEKQLACKLRLVGRPPLPVPVAVGIGTVHLHQHEDILAKT
FA
```

Do We Have a Whole Protein Domain

NCBI

CD-Search

Entrez ?

Search the [Conserved Domain Database](#) with Reverse Position Specific BLAST

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Search Database:

Enter query as **Protein**

```
MMDGIKLDWATAEALALGSLLAQGFNVRLSGQDVGRGTFSSQRHAIIVVCQETDDTYIPLNH
MDPNQKGFLEVSNSPLSEEAVLGFEYGMSIESPKLLPLWEAQFGDFFNGAQIIFDTFISG
GEAKWLLQSGIVILLPHGYDGAGPDHSSCRIERFLQAAVSTLQEMAPGTTTFNPVIGDSSV
DPKKVKTLVFCSGKHFYSLVKQRESLGAKKHDFAIIRVEELCPFPLDSLQEQEMSKYKHVK
DHIWSQEEPQNMGPWSFVSPRFEKQLACKLRLVGRPPLPVPVAVGIGTVHLHQHEDILAKT
FA
```

Please read about [FASTA](#) format description

An E1 Dehydrogenase Domain



NEW other proteins containing these domains

● .. This CD alignment includes 3D structure. To display structure, download [Cn3D v3.00!](#)

PSSMs producing significant alignments:

		Score (bits)	E value
●	gnl Pfam pfam00676 E1_dehydrog, Dehydrogenase E1 component	97.4	3e-21
●	gnl Pfam pfam01331 mRNA_cap_enzyme, mRNA capping enzyme	35.8	0.010

● [gnl|Pfam|pfam00676](#), E1_dehydrog, Dehydrogenase E1 component.

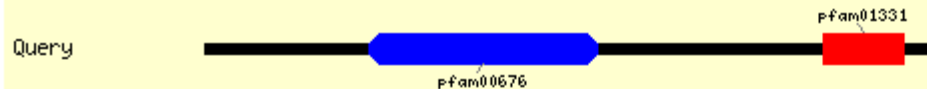
query to multiple alignment, display sequences

CD-Length = 301 residues, 89.4% aligned
Score = 97.4 bits (241), Expect = 3e-21

Query: 195 AESMMGFFH-----ELLKMSAYSGI--TDV IIGMPH RGRLLNLLTGLLQFPPELMFRKMKG 247
Sbjct: 19 RQRRRGFYHLYAGQEAALQVGIAALNPGDYI IPT-YRDHGFLLARGV--SLEEVFAELYG 75

Query: 248 LSEFPENFSATGDVLSHLTSSVDLYFGAHHPLHVTMLPNPSHLEAVNPVAVGKTRGRQQS 307

DART Results



Similar domain architectures

T50617
Homo sapiens
hypothetical prote



RPS>>

AAF26472
A. thaliana
T25K16.8 [Arabidop



RPS>>

2
Sequences
transcriptional re



T08140
C. reinhardtii
1-deoxy-D-xylulose



RPS>>

10
Sequences
PROBABLE 1-DEOXY-D



4
Sequences
capping enzyme 1a



15
Sequences
mRNA capping enzym



5
Sequences
2-oxoglutarate dehy



203
Sequences
oxoglutarate dehyd



Visualizing the Domain Hit in 3D

NCBI CD-Browser Entrez ?

CD: [pfam00676](#), CD-Search result with query-sequence added

Description: E1_dehydrog, Dehydrogenase E1 component.

Source: [Pfam\[US\]](#), [Pfam\[UK\]](#)

• This CD alignment includes 3D structure. To display structure, download [Cn3D v3.0](#)!

Redisplay Alignment showing sequences

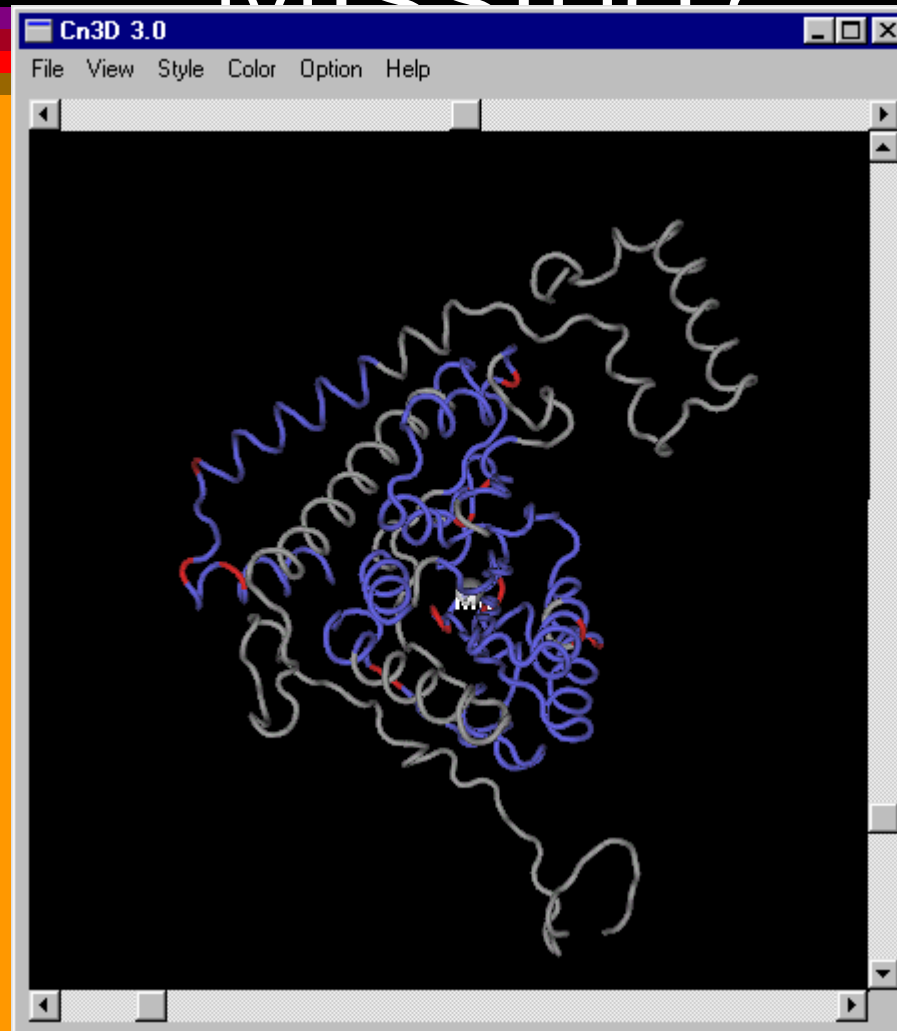
☒ Aligned chains ☒ Virtual Bonds ☒ Launch Cn3D ☐ FASTA with gaps ☐ Conservation color threshold

☐ All chains ☐ All Atoms ☐ HTML Display ☐ Phylip format ☐ Text Display

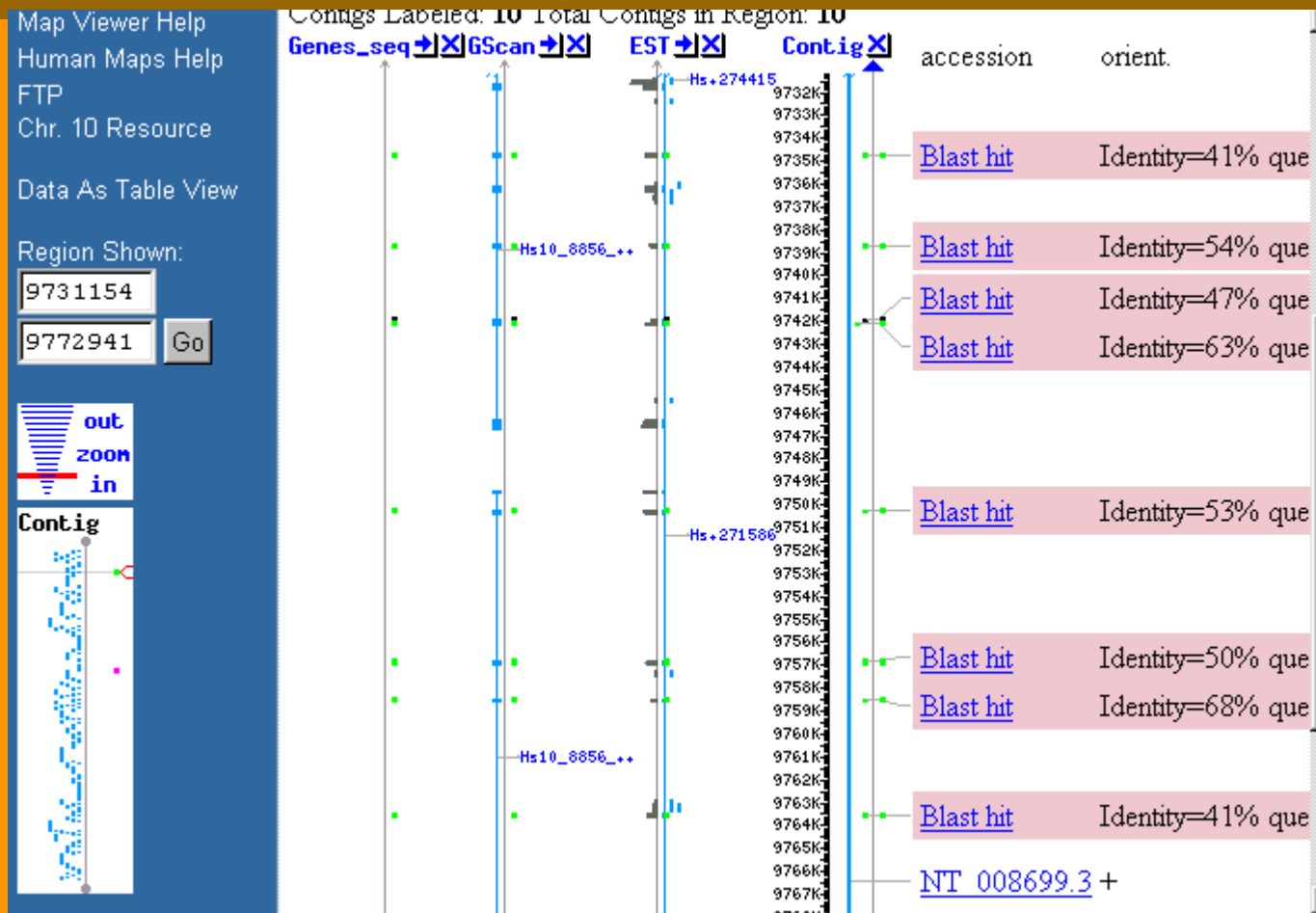
Alignment width:

		10	20	30	40	50	60	
consensus	1	YRMTLRRMEDARDALYQ	RQGRRGFYHLYAGQ	EALQVGIAAALNPGDYIPT	YRDHGFL	59		
query	177	dhflatkfistvkryggag	AESMMGFFH----	ELLKMSAYSIGI--	TDVIIGMphRGRLNL	229		
1DTW_A	61	YKSMTLLNTMDRILYESQ	RQGRISFYMTNYGEE	EGTHVGSAAALDNTDLVFGQ	YREAGVL	119		
gi 129063	66	RMMQTVRRMELKADQLYK	QKIIRGFCHLCDGQE	EACCVGLEAGINPTDHLITA	YRAHGFT	124		
gi 1709451	83	EKMVTIRRLELACDALYK	AKKIRGFCHLSIGQE	EAVAAGIEGAILDSDSIITS	YRCHGFA	141		
gi 730222	86	KDMVIIRRMEMACDALYK	AKKIRGFCHLSVGQE	EIAVGIENAITKLDSDSIITS	YRCHGFT	144		
gi 1709450	88	EDMLLCMEEDMCAQWY	HCIMRCEIHLVAGQ	EALQVGIAAALNPGDYIPT	YRDHGFL	59		

Cn3D View of the Domain Hit: What is Missing?

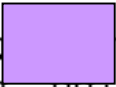


Did We Miss any Exons?



The Sequence of the Missing Exon

```
Score =
53.5 bits (127), Expect = 7e-05
Identities = 34/81 (41%), Positives = 45/81 (54%)
Frame = +2

Query: 789   EHSSARPERFLQMCNDDPDVLPDLKEANFDINQLYDCN  CSTPGNFFHVLRRQILL 848
            EHSS          QMC          D E D + + N VV+ +TP +FH+LRRQ++
Sbjct: 47207 EHSSPFLPLLSQMC-----DSAEEGVDGDTV---NMFVVHPTTPAQYFHLLRRQMVR 47353

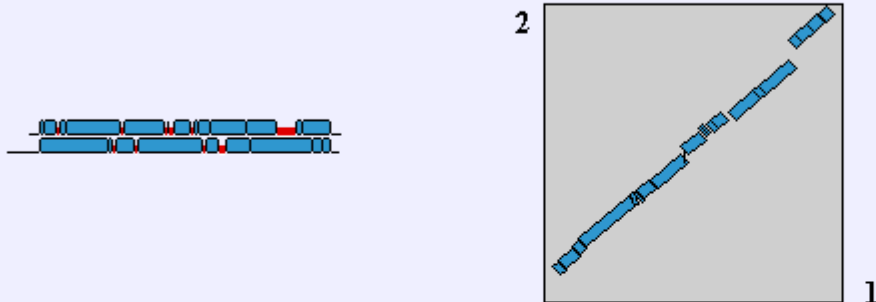
Query: 849   PFRKPLIIFTPKSLLRHPEAR 869
            FRKPLI+ +PK LLR P +R
Sbjct: 47354 NFRKPLIVASPKMLLRLLPVSR 47416

Score =
```

GenomeScan Protein vs the Probe

x_dropoff: expect: wordsize: [Filter](#) ☒ [Align](#)

Sequence 1 lc|seq_1 Length 860 (1 .. 860)
Sequence 2 gi|4505493 Length 1002 (1 .. 1002)




2

1

NOTE: The statistics (bitscore and expect value) is calculated based on the size of nr database

Score = 443 bits (1140), Expect = e-123
Identities = 309/916 (33%), Positives = 448/916 (48%), Gaps = 184/916 (20%)



Query: 42 EPQGALERPPVDH-GLARLVTVYCEHGHKAAKINPLFTGQALLENVF
E Q +++ DH + L+ Y GH A+++PL A L++
Sbjct: 115 EAQPNVDKLVEDHLAVQSLIRAYQIRGHHVAQLDPLGILDADLDSSV
oxoglutarate dehydrogenase (lin) 115 *****

The Gap is Here

```
580 A QGFNVRLSGQDVGRGTF SQRHAI VVCQETDD-TYIPLNHMDPNQKGFLEVSNSPLSEEA 638
      +G ++RLSGQDV RGTFS RH ++ Q D T IP+NH+ PNQ + V NS LSE
666 KEGIHIRLSGQDVERGTF SHRHHLHDQNVDKRTCIPMNLWPNQAPYT-VCNSSLSEYG 724
hydrogenase (lip> 666 *****

639 VLGFEYGM SIESPKLLPLWEAQFGDFFNGAQIIFDTFISGGEAKWLLQSGIVILLPHGYD 698
      VLGFE G+ + SP L LWEAQFGDF N AQ I D FI G+AKW+ Q+GIV+LLPHG +
725 VLGFEAGLRMASPNALVLWEAQFGDFHNTAQCIIDQFICPGQAKWVRQNGIVLLLPHGME 784
hydrogenase (lip> 725 *****

699 GAGPDHSSC RIERFLQ----- 714
      G GP+HSS R ERFLQ
785 GMGPEHSSARPERFLQMCND DDPVLPDLKEANFDINQLYDCNWWVVNCSTPGNFFHVLRR 844
hydrogenase (lip> 785 *****

715 -----AAVSTLQEMAPGTT FNPVI---GDSSVDPKKVKT LVFC 749
      A S+ EM PGT F VI G ++ +P+ VK L+FC
845 QILLPFRKPLIIFTPKSLLRHPEARSSFDEMLPGTHFQRVIPEDGPAAQNPENVKRL LFC 904
hydrogenase (lip> 845 *****

750 SGKHFYSLVKQRESLGAKKHDFAIIRVEELCPFPLDSLQQEMSKYKHVKDHIWSQEEPQN 809
      +GK +Y L ++R++ AI R+E+L PFP D L +E+ KY + + W QEE +N
905 TGKVVYYDLTRERKARD-MVGQVAITRIEQLSPFPFDLLLKEVQKYPNA-ELAWCQEEHKN 962
hydrogenase (lip> 905 *****

810 MGPWSFVSPRFEKQLA 825
      G + +V PR ++
963 QGYDYVVKPRLRTTIS 978
hydrogenase (lip> 963 *****
```

Finding a Structural Template for a Protein Sequence

Visualizing the
Superposition of Sequence
on Structure with Cn3D

GenScan Predicted Peptide and ScanProsite Results

>14:10:32|GENSCAN_predicted_peptide_2|424_aa

MSQICKRGLLISNRLAPAALRCKSTWFSEVQMGPPDAILGVTEAFKKDTNPKKINLGAGA
YRDDNTQPFVLPVREAEKRVVSRSLDKEYATIIGIPEFYNKAIELALGKGSKRLAAKH
VTAQSISGTGALRIGAAFLAKFWQGNREIYIPSPSWGNIHVAIFEHAGLPVNRYYDKDT
CALDFGGLIEDLKKIPEKSIVLLHACAHNPTGVDPTLEQWREISALVKKRNLYPFIDMAY
QGFATGDIDRDAQAVRTFEADGHDFCLAQ**SFAKNMGLYGERAG**AFTVLCSDDEEEAARVMS
QVKILIRGLYSNPPVHGARIAAEILNNEDLRAQWLKDVKLMDRIIDVVRTKLKDNLIKLG
SSQNWDHIVNQIGMFCFTGLKPEQVQKLIKDHVYLTNDGRVSMAGVTSKNVEYLAESIHK
VTK

[GS]-[LIVMFYTAC]-[GSTA]-**K**-x(2)-[GSALVN]-[LIVMFA]-x-[GNAR]-
x-R-[LIVMA]-[GA] [**K is the pyridoxal-P attachment site**]

270-283 SFAK**NMGLYGERAG**

Begin with a CDD Search

NCBI

CD-Search

Entrez ?

Search the [Conserved Domain Database](#) with Reverse Position Specific BLAST

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Search Database:

Enter query as **Protein**

```
CALDFGGLIEDLKKIPEKSI VLLHACAHNPTGVDPTLEQWREISALVKKRNL YPFIDMAY
QGFATGDIRDAQAVRTFEADGHDFCLAQSFAKNMGLYGERAGAF TVLCSDEEEAARVMS
QVKILIRGLYSNPPVHGARIAAEILNNEDLRAQWLKDVKLMADRIIDVRTKLKDNLIKLG
SSQNWDHIVNQIGMFCFTGLKPEQVQKLIKDHVS YLTNDGRVSMAGVTSKNVEYLAESI H
KVTK
|
```

Please read about [FASTA](#) format description

A Domain Hit...

gnl|Pfam|pfam00155, aminotran_1, Aminotransferase class-I

Add query to multiple alignment, display sequences

CD-Length = 404 residues, 99.8% aligned

Score = 410 bits (1055), Expect = 5e-116

Query:	23	KSTWFSEVQMGPPDAILGVTEAFKKTNPCKINLGAGAYRDDNTQPFVLPVREAEKRVV	82
Sbjct:	1	LSSMFVRVSHAPGDPILGVWEAFKEDPRPGKINIGLGIYEPDLGKDLVLPVAVKKAEARLA	60
Query:	83	-SRSLDKEYATIIGIPEFYNKAIELALGKGSKRLLAAKHNVTAQSIGTGALRIGAAFLAK	141
Sbjct:	61	LDRGGFKEYLP IHGLPEFREAIKLYFGDRSPALKFKRVEVVQTLGGTGALRLAADFLAN	120
Query:	142	FWQGNREIYIPSPSWGNHVAIFEHAGLPVNRVRYD KDTCA LDFGGLIEDLKKIPEKSI V	201
Sbjct:	121	P---GDEVLPDPTWPNYADIFKAAGFEVIPYRYDENNFKLD FEAL EAAITEAPEKTKV	177
Query:	202	LLHACAHNPTGVDPTLEQWREISALVKKRNLYPFIDMAYQGFATGDI DRDAQAVRTFEAD	261
Sbjct:	178	LLHNNPHNPTGTDPTREQLKKIAAVVKEKNILLSSDEAYQGFVFGDL--DAASVAEFAEE	235
Query:	262	GHD FCLAQSF AKNMGLYGERAG AFTVLCSD E-----E AARVMSQVKILIRGLYSNPPV	315
Sbjct:	236	GDELLVVQSFSKNFGLYGWRVGAIVVCAIINAAAKKSSAGRVSSQLQSLARAMYSNPPD	295
Query:	316	HGARIAAEILNNEDLRAQWLKDVKLMADRIIDVRTLKDKNLIKLGSSQNWDHIVNQIGMF	375
Sbjct:	296	HGA EIVARILSRPDLFTSWLEEVKGMACRIPNGRFYLWPDLSKLGRPE--DHIFEQDGMF	353
Query:	376	CFTGLKPEQVQ-KLIK DHSVYLTNDGRVSMAGVTSKNV EYLAESIHKVTK	424
Sbjct:	354	SFTLLEEAQVVVIPGSEFGIYEPGWGRISLAGLSEANVDEAAERIRAFVK	403

A Template Emerges...

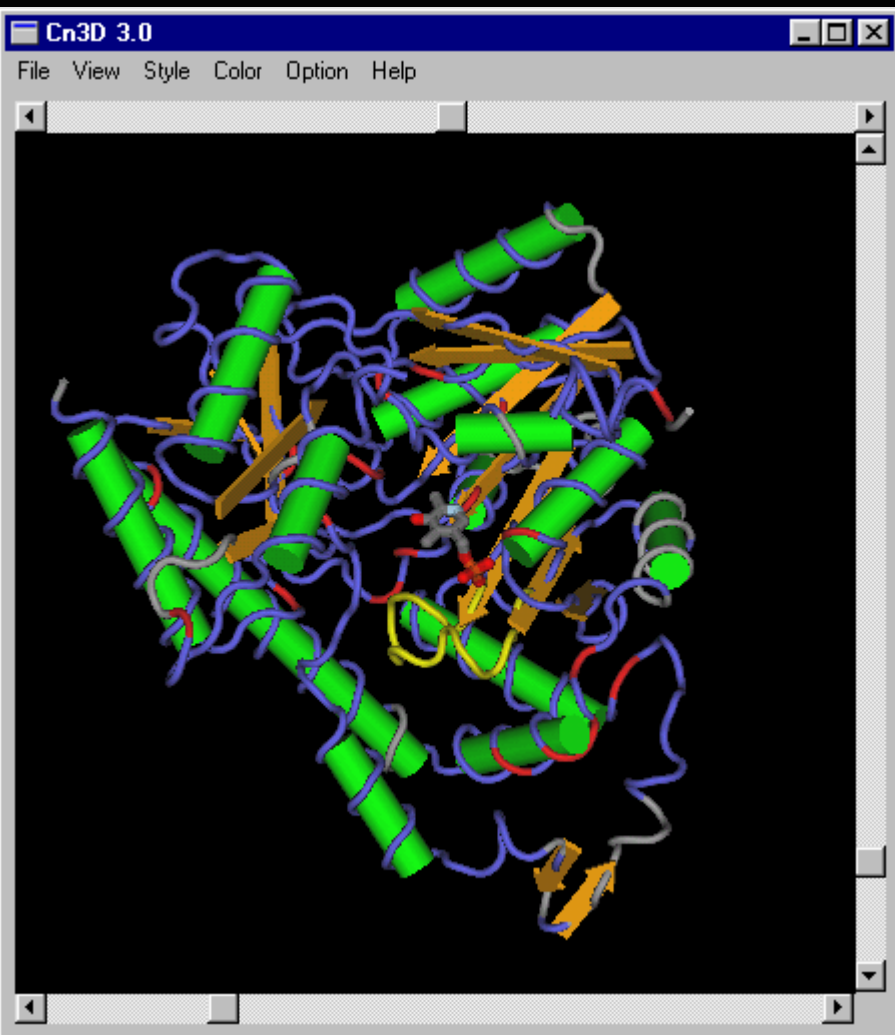
Redisplay Alignment showing up to 10 sequences most similar to the query

☒ Aligned chains ☒ Virtual Bonds ☒ Launch Cn3D ☐ FASTA with gaps ☐ Conservation color threshold
☐ All chains ☐ All Atoms ☐ HTML Display ☐ Phylip format ☐ Text Display Alignment width: 60

		10	20	30	40	50	60	
	******	
consensus	1	LSSMFVRVSHAPGDP	ILGVWEAFKEDPRPGKIN	----	IGLGIYEPDLGKDLVLP	PAVKKAE	56	
query	23	KSTWFSEVQMGPDA	ILGVTEAFKKDTNPKKIN	----	LGAGAYRDDNTQPFV	LPSVREAE	78	
1B8G B	2	LSRNATFNSHGQDSS	YFLGWQEYEKNPYHEVHN	tngi	IQMGLAENQLCFDL	LLESWLAKNP	61	
7AAT A	1	-SSWWSHVEMGPPDP	ILGVTEAFKRDTNSKKMN	----	LGVGAYRDDNGKPY	VLNCVRKAE	55	
gi 1168256	29	MSSWWKSVEPAPKDP	ILGVTEAFLADPSPEKVN	----	VGVGAYRDDNGKPV	VLECVREAE	84	

		70	80	90	100	110	120	
	******	
consensus	57	ARLALDRGG	-----	FKEYLPIHGLPEF	REAI	AKLYFGDRSPALK	FKRVEV	VQTLGGTGA 110
query	79	KRVV-SRSL	-----	DKEYATIIGIPEF	YNKA	IELALGKGSKR	LAAKHNV	TAQSIGTGA 131
1B8G B	62	EAAAFKKNGesifae	LALFQDYHGLPA	FKKAM	VDFMAEIRGNKV	TDPNHLVLT	AGATSA	121
7AAT A	56	AMIAAKK-M	-----	DKEYLPIAGLADF	TRASA	ELALGENSEAF	KSGRYV	TVQGISGTGS 108
gi 1168256	85	KRLAGS--T	-----	FMEYLPMGGS	AKMVDL	TLKLAYGDNSE	FIKDKRIA	AVQTLSGTGA 136

The Prosite Motif Unmasked

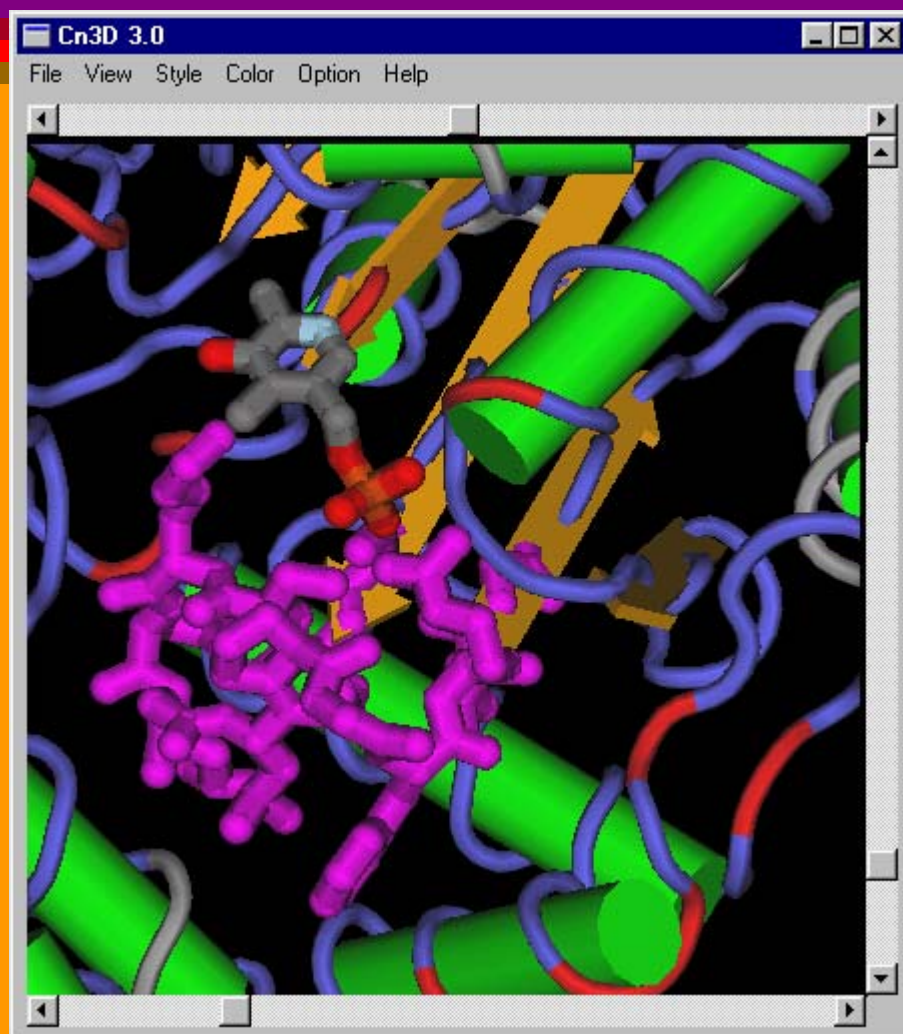
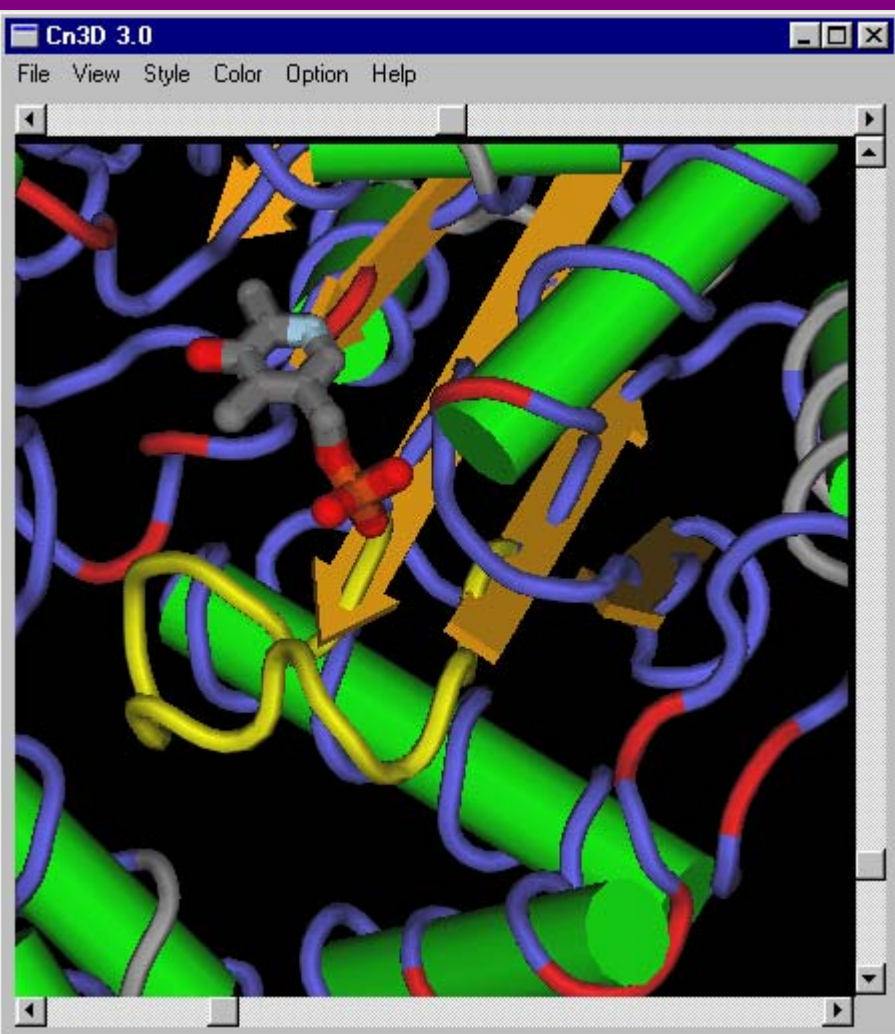


s Help

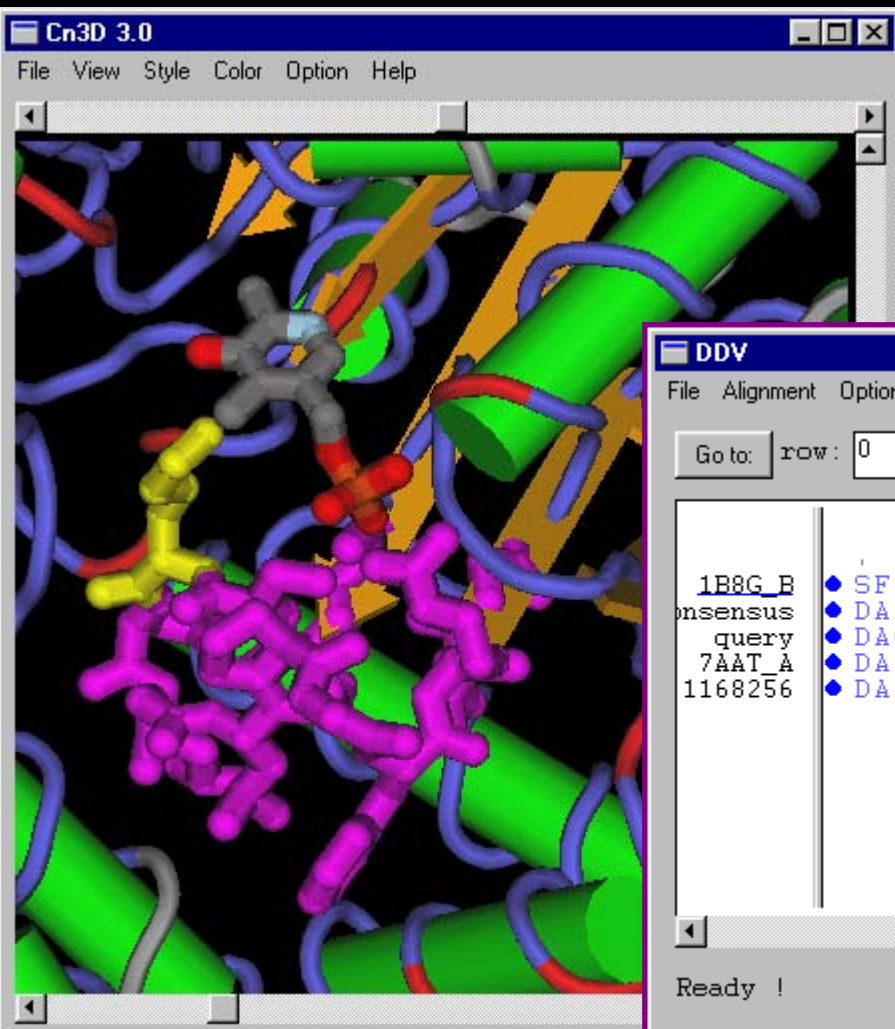
col: 0

280	290	300	310	320
ISVMEVLKD	rnndensev	WQRVHVVY	SLSKDLGLPGFRVG	AIYSNDd
ASVAEFAEE	~~~~~	GDELLVVQ	SFSKNFGLYGWRV	GAIIVVVCa
QAVRTFEAD	~~~~~	GHDFCLAQ	SFAKNMGLYGERA	GAFVLCs
VALRHFIEQ	~~~~~	GIDVVLISQ	SYAKNMGLYGERA	GAFVVIC~
KSIRIFLED	~~~~~	GHIGISQ	SYAKNMGLYGRV	GCISVLC~

Close-ups of the Motif



The Critical Lysine



DDV

File Alignment Options Help

Go to: row: 0 col: 0

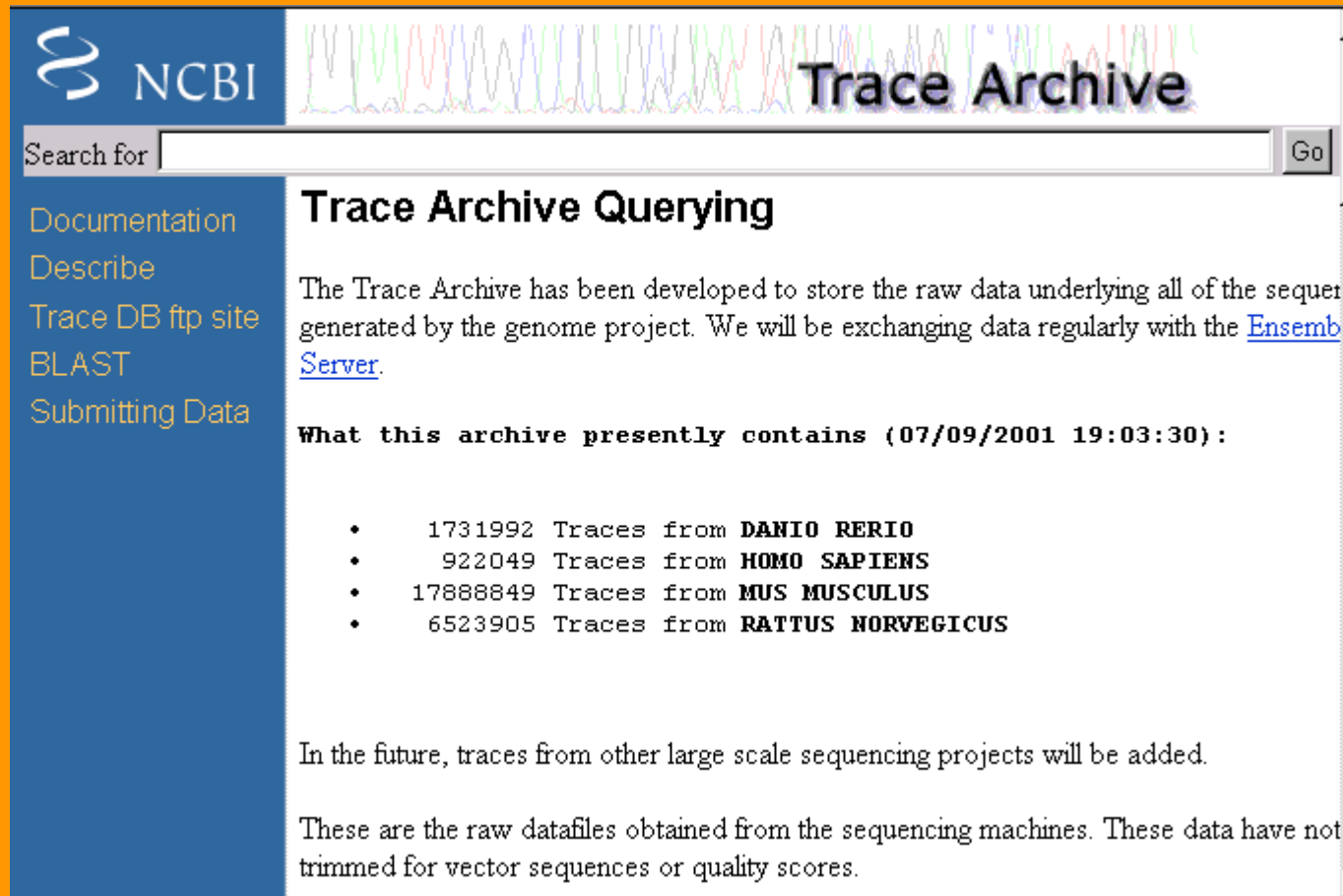
		280		290		300		310		320	
1B8G_B	◆	SFISVMEVLKD	rncdensev	WQRVHVVY	SLS	DLGLPGFRV	GAIYSND	d			
onsensus	◆	DAASVAEFAEE	~~~~~	GDELLVVQ	SFS	KNFG	LYGWR	VGAIVVVC	a		
query	◆	DAQAVRTFEAD	~~~~~	GHDFCLAQ	SFA	KNMGLY	GERAGA	FTVLC	s		
7AAT_A	◆	DAWALRHFIEQ	~~~~~	GIDVVL	SQSYA	KNMGLY	GERAGA	FTVIC	~		
1168256	◆	DAKSIRIFLED	~~~~~	GHHIGISQ	SYA	KNMGLY	GQRV	GC	SVLC	~	

Ready !

Piecing Together a Gene from the Mouse Trace Archive

NX-57: A Mouse Kidney-
Specific Membrand Protein

The NCBI Trace Archives



The screenshot shows the NCBI Trace Archive website. At the top left is the NCBI logo. To its right is a decorative header with a colorful DNA sequence trace pattern and the text "Trace Archive". Below the logo is a search bar with the placeholder text "Search for" and a "Go" button. A left-hand navigation menu contains links: "Documentation", "Describe", "Trace DB ftp site", "BLAST", and "Submitting Data". The main content area is titled "Trace Archive Querying" and contains the following text:

The Trace Archive has been developed to store the raw data underlying all of the sequen generated by the genome project. We will be exchanging data regularly with the [Ensemb Server](#).


What this archive presently contains (07/09/2001 19:03:30):

- 1731992 Traces from **DANIO RERIO**
- 922049 Traces from **HOMO SAPIENS**
- 17888849 Traces from **MUS MUSCULUS**
- 6523905 Traces from **RATTUS NORVEGICUS**

In the future, traces from other large scale sequencing projects will be added.

These are the raw datafiles obtained from the sequencing machines. These data have not trimmed for vector sequences or quality scores.

BLAST Against WGS Reads

 **NCBI**

megablast **BLAST**

Nucleotide Protein Translations Retrieve results for an RID


Trace Archive database Mega BLAST search

Search

MM_020626

Load query file from disk

[Set subsequence](#) From: To:

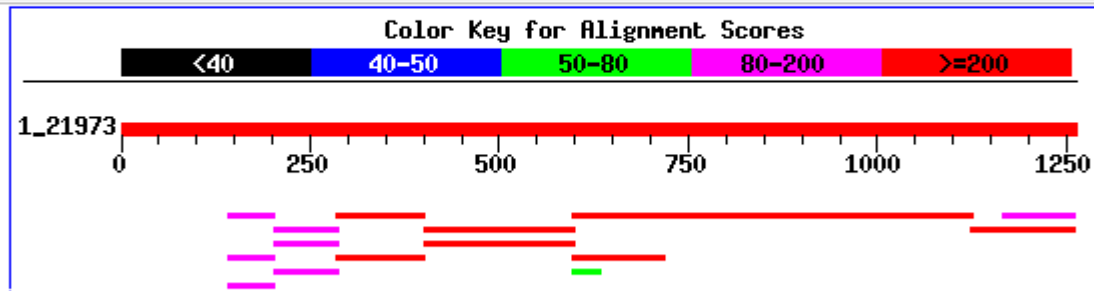
[Database](#) 

[Return alignment endpoints only](#) ☐

The Coverage is Good, but None of the Reads Completely Spans an Intron

Distribution of 15 Blast Hits on the Query Sequence


Mouse-over to show defline and scores. Click to show alignments



Genomic Comparisons

COGs & Taxplot

K12 vs O157 COGwise

<i>Escherichia coli</i> k12	-	x	-	x	-	x	-	x	 E - Enterobacteriaceae
<i>Escherichia coli</i> O157	-	-	x	x	-	-	x	x	
<i>Buchnera sp.</i> APS	-	-	-	-	x	x	x	x	
Function	1219	50	64	1281	2	2	-	548	
J K L	187	9	9	141	-	-	-	172	Information storage and processing
J	78	-	-	19	-	-	-	116	Translation, ribosomal structure and biogenesis
K	49	3	4	52	-	-	-	16	Transcription
L	60	6	5	70	-	-	-	40	DNA replication, recombination and repair
D O M N P T	161	9	11	329	-	1	-	129	Cellular processes
D	6	-	-	15	-	-	-	10	Cell division and chromosome partitioning
O	27	-	3	46	-	-	-	30	Posttranslational modification, protein turnover, chaperones
M	32	3	1	85	-	-	-	27	Cell envelope biogenesis, outer membrane

Some Differences...

9 [COGs](#)

Protein/Gene name:

[Select](#)

[Help](#)

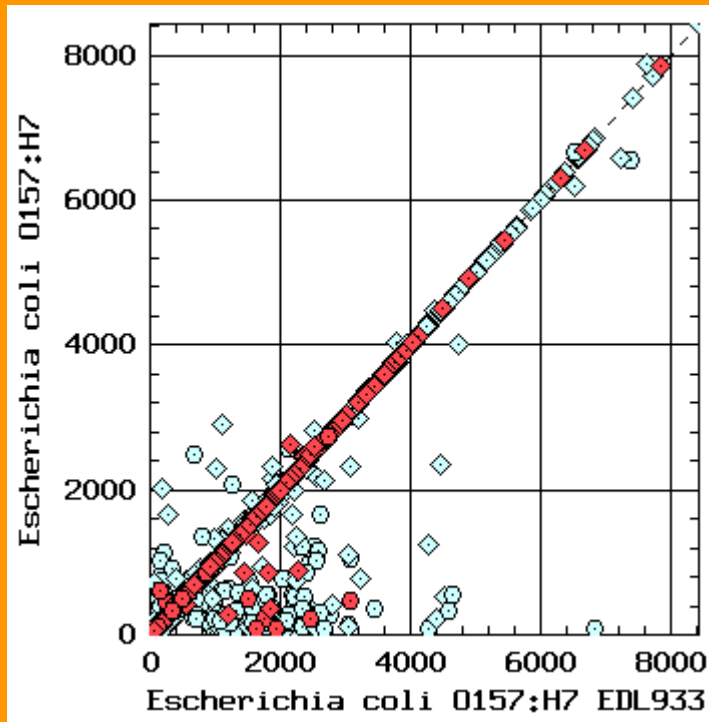
[Functional](#) categories:

- D Cell division and chromosome partitioning
- O Posttranslational modification, protein turnover, chaperones
- M Cell envelope biogenesis, outer membrane
- N Cell motility and secretion
- P Inorganic ion transport and metabolism
- T Signal transduction mechanisms

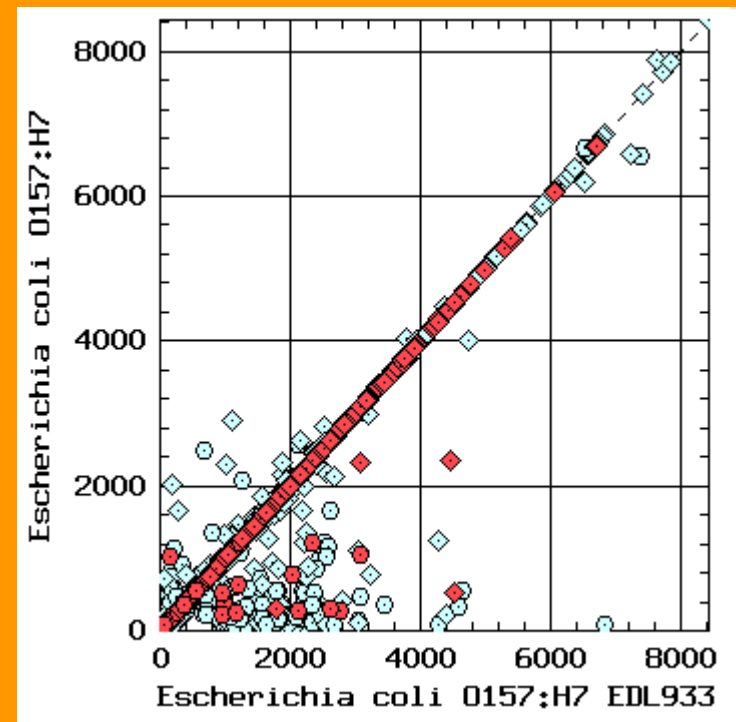
Only *Escherichia coli* k12

-	7	-----qv---b-efg-----	FlgM	[KN]	COG2747	Negative regulator of flagelli
-	8	--m-----dr---e-----u---w	Glf	[M]	COG0562	UDP-galactopyranose mutase
1	28	a-mpkz---drlbcef-hsnuj----	RfbC	[M]	COG1898	dTDP-4-dehydrorhamnose 3,5-epi
-	26	aompkz-q-drlbcef--sn-jx----	RfbD	[M]	COG1091	dTDP-4-dehydrorhamnose reducta
-	3	-----efg-----	Flh0	[N]	COG3190	Flagellar biogenesis protein
-	4	-----q-----e-g-----	GspC	[N]	COG3031	General secretion pathway prot
-	6	-----efg-s--j-----	GspK	[N]	COG3156	General secretion pathway prot
-	3	-----efg-----	EtpM	[N]	COG3149	General secretion pathway prot
-	5	-----efg---j-----	GspL	[N]	COG3297	General secretory pathway prot

Divergence from K12 Seen Using the TaxPlot

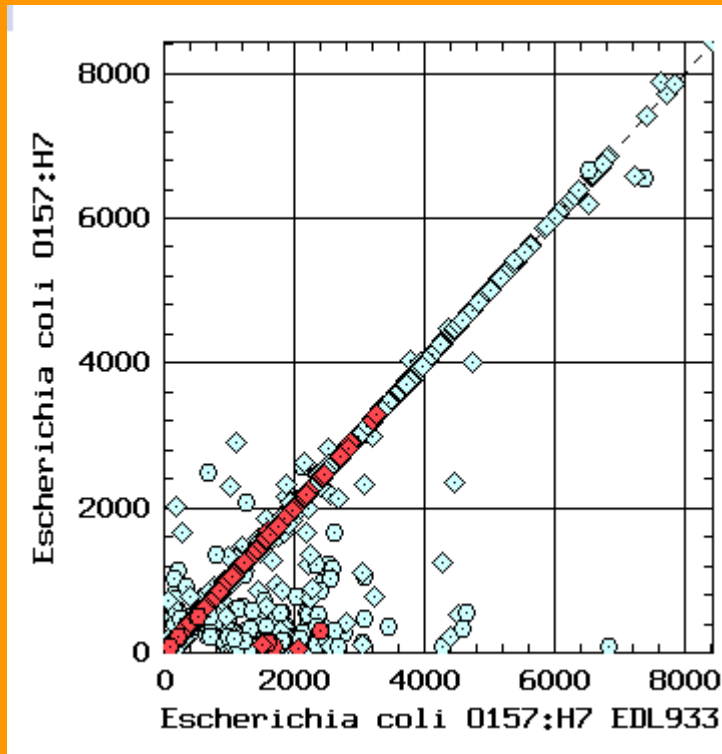


Amino Acid Transport & Metabolism

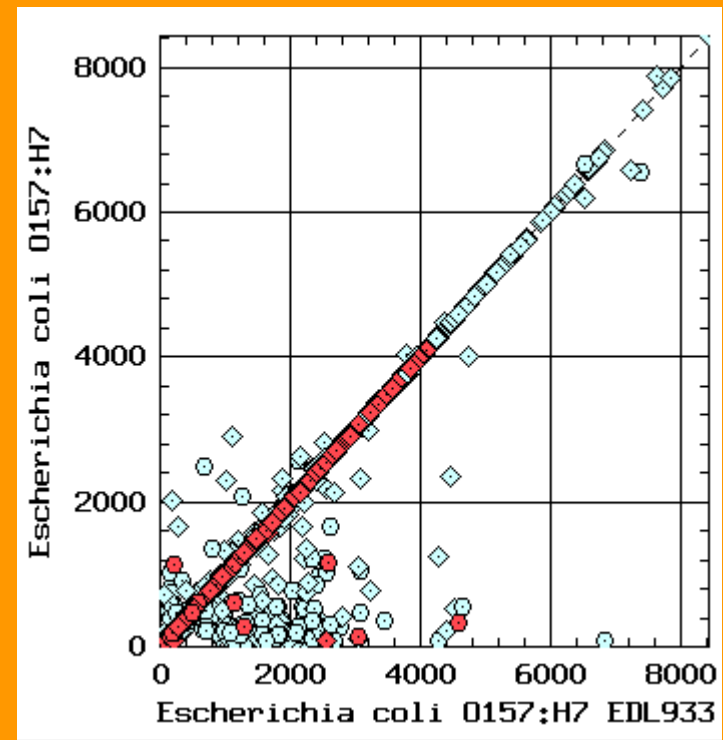


Energy Production

Divergence from K12



Coenzyme Metabolism



Inorganic Ion Transport

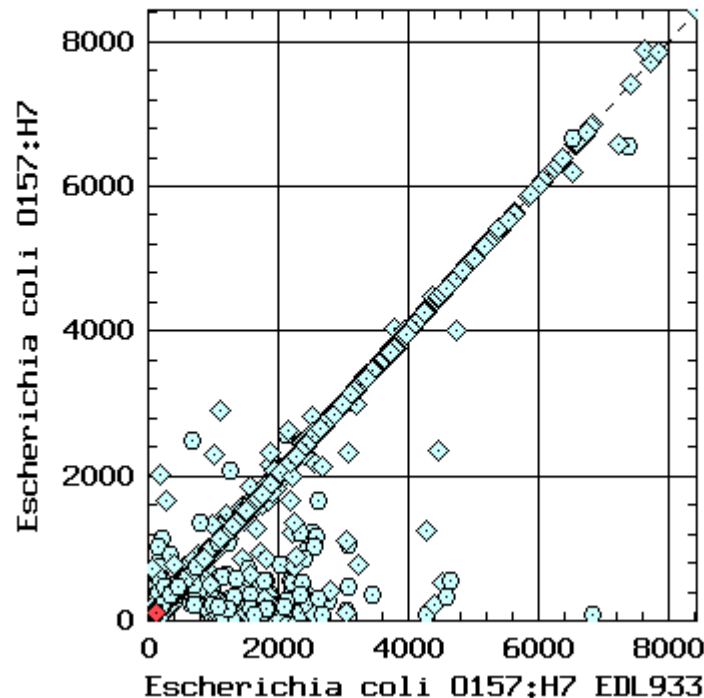
YadC: The Most Divergent Gene

Distribution of *Escherichia coli* K12 homologs

596 hits

2755 equal hits

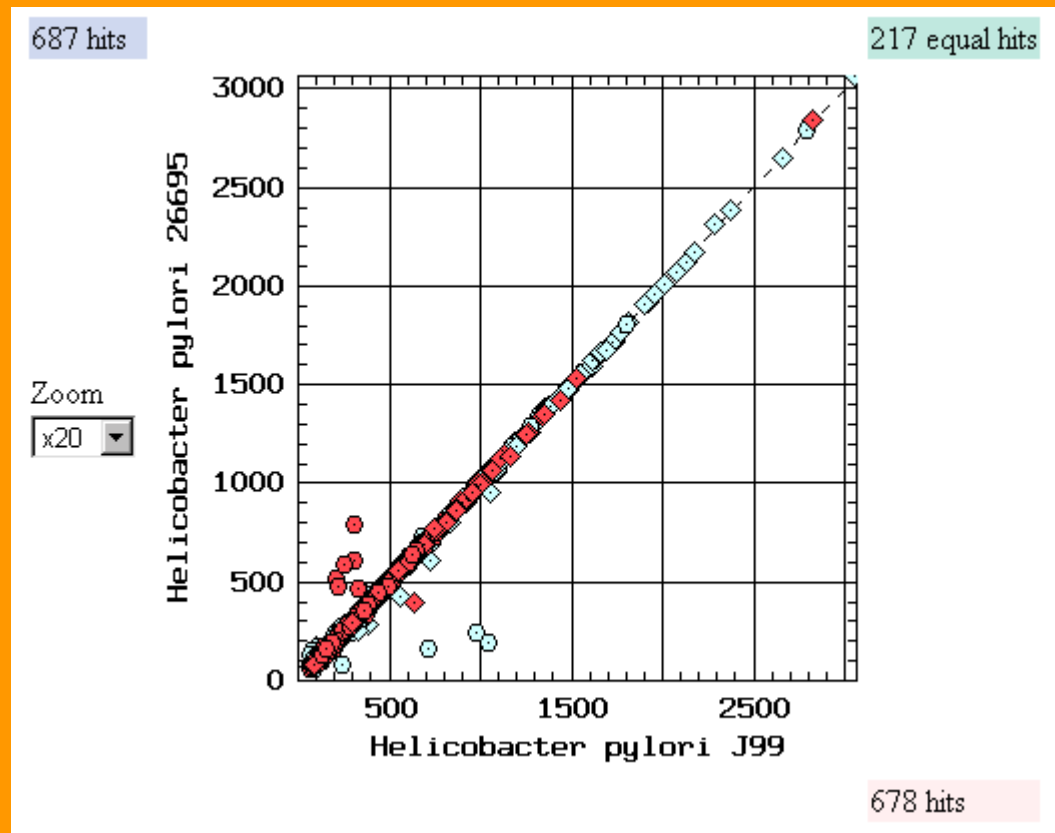
Zoom
x20



The Two Pylori's COGwise

<i>Helicobacter pylori</i> 26695	-	x	-	x	-	x	-	x	<div>U - Proteobacteria epsilon subdivision</div> 
<i>Helicobacter pylori</i> J99	-	-	x	x	-	-	x	x	
<i>Campylobacter jejuni</i>	-	-	-	-	x	x	x	x	
Function	2004	4	7	100	258	3	2	788	
J K L	278	1	-	14	22	-	-	203	Information storage and processing
J	92	-	-	1	6	-	-	114	Translation, ribosomal structure and biogenesis
K	84	-	-	3	10	-	-	27	Transcription
L	102	1	-	10	6	-	-	62	DNA replication, recombination and repair
D O M N P T	328	1	1	25	62	2	1	220	Cellular processes
D	15	-	-	4	-	-	-	12	Cell division and chromosome partitioning
O	49	-	-	5	5	-	-	47	Posttranslational modification, protein turnover, chaperones
M	65	-	-	6	14	2	-	61	Cell envelope biogenesis, outer membrane

K12 vs H. pylori Using the TaxPlot



Amino Acid Transport and Metabolism